DOCUMENT RESUME

ED 422 362 TM 028 919

AUTHOR Holweger, Nancy; Weston, Timothy
TITLE Differential Item Functioning: An Applied Comparison of the Item Characteristic Curve Method with the Logistic Discriminant Function Method.
PUB DATE 1998-00-00
NOTE 43p.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Item Bias; Item Response Theory; Performance Based Assessment; State Programs; *Test Items; Testing Programs
IDENTIFIERS Item Bias Detection; Item Characteristic Function; *Logistic Discriminant Function Analysis

ABSTRACT

This study compares logistic discriminant function analysis for differential item functioning (DIF) with a technique for the detection of DIF that is based on item response theory rather than the Mantel-Haenszel procedure. In this study, the areas between the two item characteristic curves, also called the item characteristic curve method is compared with the logistic discriminant function analysis method for each item and the entire test. Data from a state-level performance-based assessment program for approximately 16,000 examinees are used. Substantial differences are found between the items identified as having DIF using the item characteristic curve method and the logistic discriminant function method. Possible reasons for these differences are discussed. Since a clear determination about the best method to determine DIF is not apparent, manufactured data should be used in a Monte Carlo study to eliminate design aspects of the assessment that confused these results. Appendixes contain plots of item characteristic curves and illustrative figures. (Contains 7 plots, 7 figures, and 13 references.)(SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 422 362

Differential Item Functioning: An Applied Comparison of the Item Characteristic Curve Method with the Logistic Discriminant Function Method

by

Nancy Holweger, University of Colorado

and

Timothy Weston, University of Colorado

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Nancy Holweger

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM028919

Theoretical Framework

The fact that certain items in a test may be biased against certain groups has become a matter of considerable concern to the examinees, the users of tests, and the testing community (Berk, 1982). There has been little agreement as to the definition of item and test bias and, consequently, the techniques for detecting bias have been limited (Hambleton & Swaminathan, 1985). There are numerous procedures that are used to detect item bias (Camilli & Shepard, 1994). Those based upon item response theory fall into one of the following three categories:

1. Comparison of item characteristic curves.
2. Comparison of the vectors of item parameters.
3. Comparison of the fit of the item response models to the data.

These procedures have been developed primarily for use with dichotomously scored, multiple-choice test items.

The recent emphasis on performance assessment has created the need for techniques that can accommodate item formats other than those scored correct/incorrect. These assessments encompass item responses that are scored on a nominal or ordinal scale involving more than two categories (Miller & Spray, 1993). Such polytomous items are typically treated as a series of dichotomous pairs when using the item response procedures for differential item functioning (DIF) that have been listed above. For example, response A is either correct or incorrect. This can be determined for each of the item choices (such as B, C, etc.) or for different levels of achievement (such as Emergent, Meets Expectations, Mastery, etc.). Each of these artificially dichotomous options contributes to the DIF of a particular item. Clearly, simultaneous estimation of all item choices is desirable.

Logistic regression has been used successfully to detect instances of uniform and nonuniform DIF in simulations of dichotomous item responses (Swaminathan & Rogers, 1990). This procedure yields the probability of observing each dichotomous

item response as a function of the observed test scores and a group indicator variable. Several logistic regression techniques for calculating DIF have the same limitations as do those using item response theory. These include continuation-ratio logits and cumulative logit models. However, according to Miller and Spray (1993), logistic discriminant function analysis circumvents this limitation because it allows the response variable to assume any number of categories, and it can take on any one of the values associated with each item. These authors demonstrate the application of this technique to a mathematics performance test developed by American College Testing and compare it with the widely used Mantel-Haenszel procedure for uniform DIF. They obtain similar results in these two determinations.

Purpose

This study compares logistic discriminant function analysis for differential item functioning (DIF) with a technique for the detection of DIF that is based upon item response theory rather than the Mantel-Haenszel procedure. In this study, the areas between the two item characteristic curves, also called the item characteristic curve method is compared with the logistic discriminant function analysis method for each item and the entire test. The question of concern in this study is as follows:
How does the logistic discriminant function analysis for differential item functioning (DIF) compare to the item characteristic curve method for DIF determination on a major science performance assessment?

Data Source

Sample

Test data from a recently administered state-level performance-based assessment program is used for these analyses. The sample for analysis includes approximately 16,000 examinees. After removing students with total zero scores, our sample

consists of 8,539 females and 8,029 males. We remove students who score a zero on all items because the program we use to calculate the ICC method of DIF, Multilog (SSI, 1991), does not function with total zero scores.

Structure of the Assessment

The tasks comprising this assessment are an integrated series of complex performance tasks designed to elicit higher-order thinking skills. The tasks are grouped in four to six test booklets and a matrix sample is used to collect a variety of school level information. Items draw from the following domains:

1. Concepts of Science: This area deals with unifying themes from life, physical, earth and space science.
2. Nature of Science: This area asks students to explain and interpret information about scientific phenomena.
3. Habits of the Mind: This area asks students to demonstrate ways of thinking about science.
4. Attitudes: Not assessed.
5. Science Processes: This area asks students to use language, instruments, methods and materials to collect, organize and interpret information.
6. Application of Science: This area asks students to apply what they have learned in science to solve problems.

While this test is primarily a test of science knowledge, many of the items on the test are a mixture of science and reading or science and math. Listed below are the components included in each of the 30 test items.

Items with Science Only

Items #6, 7, 8, 12, 14, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30

Items with Science and Reading

Items #1, 2, 3, 4, 5

Items with Science and Math

Items #9, 10, 11, 13, 15, 16, 17

This assessment system is not designed for student-level accountability; instead, the results are used to monitor achievement at the school and district level.

Scoring and Calibration

The scoring of student responses is conducted in a manner similar to the scoring of many other forms of open-ended assessments. The scoring process includes the development of a scoring guide (rubric), extensive training of scorers that requires scorers to meet a minimum standard on a qualifying set of papers (70-80% exact agreement, depending on the type of task), and a series of quality control procedures including read-behinds, check-sets, and reader-effects reliability studies. To calibrate the item raw scores, a graded response model is used for converting the responses to these items into scale scores (Yen, et al., 1992).

Method

ICC

The first step in the comparison of these two methods for the determination of DIF is to conduct the analysis using the item characteristic curve method. This involves the determination of item parameters for the full group using Samejima's (1968, cited in Thissen, 1991) graded-response model with a random MML estimation procedure. Multiple *b* parameters are found for polytomous items. Of the thirty items on the assessment, 11 have two categories, 18 have three categories and 1 has four categories. For this reason, the items are examined using Multilog (SSI,1991), a software package designed to conduct IRT analyses of items with multiple response categories. The item parameters from the random MML procedure are then fixed and individual theta scores are estimated for each student in the full data set.

The identical estimation procedure is used to find item parameters and theta estimates for the separate male and female groups. As Camilli and Shepard (1994),

suggest, we calculate a baseline index in which we randomly split the reference group (female), into halves but then treat these groups as if they are reference and focal groups. We use this procedure because the sampling distribution for DIF values can only be interpreted in relation to other values on the same test. A baseline should only show the amount of DIF accounted for by random sampling variation allowing us to better estimate the real DIF.

A simple linear transformation is used to place the item parameters from the two groups on the same scale, using the means and standard deviations by gender calculated from the full-group analysis. Item characteristic curves for each item and each group are plotted using the scaled item parameters. The following linear transformation is used to put the item and ability parameters on comparable scales.

$\theta = c + d\theta^*$

$b = c + b^* (d)$

$a = a^* / d$

The c and d constants are calculated as follows:

$c = \theta - d\theta^*$

$d = S\theta / S\theta^*$

Where θ, $S\theta$, b and a equal the group mean, standard deviation, and b and a parameters of (either) the males or the females in the joint data set, and θ^*, $S\theta^*$, b^* and a^* equal the group mean, standard deviation, b and a parameters of (either) the males or the females in one of the separate data sets.

The probability index developed by Linn (1981, cited in Hambleton & Swaminathan, 1985) is used to find the unsigned area between the item characteristic curves of the males and the females:

$$DIF = \sum ((P_{i1}(\theta k) - P_{i2}(\theta k))^2 \Delta\theta)^{1/2}$$

LDF

The second step in this comparative analysis is to determine DIF using the logistic discriminant function method (Miller & Spray, 1993). "The significance of the logistic discriminant function analysis to DIF identification with polytomous item responses is that the discriminant function can be written as:

Prob (G|X, U) = e(1-G)(-α_0-α_1X -α_2U -α_3X*U) / 1 +e(-α_0-α_1 X -α_2U -α_3X*U) ...

and the response variable, U, need not be restricted to only two categories ... but can take on any one of the J values associated with each item" (p. 109). Using this technique, nonuniform DIF is determined from the difference between the above discriminant function and the the hierarchical model:

Prob (G|X, U) = e (1-G)(-α_0-α_1X -α_2U) / 1 +e(-α_0-α_1 X -α_2U)

Uniform DIF is determined by the difference between the hierarchical model and the null probability model:

Prob (G|X) = e (1-G)(-α_0 -α_1X) / 1 +e (-α_0 -α_1X)

(p. 110).

Each of these models are tested for significance and then compared.

Simultaneous confidence bands using the Scheffe' multiple comparison technique are then calculated (Hauck, 1983). The results of this logistic discriminant function analysis are then compared with the results of the analysis using the item characteristic curve method. The amount of DIF for each item in the subtest is compared using each method.

Results

Table 1 and Appendix A - ICCs

The findings from our ICC differential item functioning analyses of each of the thirty questions in the assessment can be found in Table 1. The range of DIF found for the b1 indices (the DIF between rating categories 0 and 1) is from .03. to .33. The range of DIF found for the b2 indices (the DIF between rating categories 1 and 2) is from .02. to

19. The DIF for the b3 index (the DIF between rating categories 2 and 3) is .05. Table 1 also includes the a and b parameters for each of the items for comparison between the two gender groups. Consideration of these parameters in conjunction with examination of the DIF indices is suggested by Camilli and Shepard (1994) to determine actual test item bias.

These DIF indices are calculated for uniform differential item functioning. It is customary to examine the Item Characteristic Curves for each of the items in the assessment for the purpose of determining whether the DIF is uniform or nonuniform. Appendix A includes the ICCs for the seven items with the largest DIF based upon this method (Items # 3, 18, 4, 25, 26, 23 and 5), roughly in that order of decreasing magnitude. These items display primarily uniform DIF. However, items #23 and #26 have slight nonuniform DIF. The indices for these two items are artificially low because we used an unsigned area measure to calculate their DIF.

Appendix B - LDFs

Appendix B contains depictions of the probability functions for each of the seven items in the assessment that show DIF due to gender using the logistic discriminant function technique. According to this technique, if group membership, either male or female, can be determined more accurately from the total score plus the item score than it can for the total score alone, then that item is said to have DIF. The absolute determination of DIF is obtained by comparing the probability for the total score with the .95 confidence intervals around the total score + item score. If the line for the total score falls outside the confidence bands, the "item plus total" functions differently than the total test score and gender DIF is said to exist. Each of the depictions in Appendix B show an item which has the total score line falling outside the confidence intervals. Items #12, 3, 4, 29, 5, 13 and 24, roughly in that order of decreasing magnitude, exhibit DIF according to the LDF method of determination. These seven items are the only items in this assessment that exhibit such absolute DIF.

In order to compare the two techniques, the seven items with absolute DIF according to the LDF technique are considered along with the seven items with the largest DIF indices according to the ICC technique. As can be seen from the lists of those items in the sections above, there is no uniformity across techniques. There is some overlap, however. Items 3, 4 and 5 are identified by both techniques as having substantial DIF. However, items 12, 29, 13, and 24 show DIF using the LDF technique but not the ICC technique. Items 18, 23, 25 and 26 show DIF using the ICC technique but not the LDF technique.

Conclusions

The substantial differences between the items that are identified as having differential item functioning using the item characteristic curve method versus the logistic discriminant function method are not anticipated. However, this situation is far from unusual. According to Camilli and Shepard (1994), studies comparing methods of determination of item bias often produce such differing results.

Considering this discrepancy in results, a discussion of the possible reasons for these differences is necessary. In addition, the comparison of these two methods, the original research question posed by this project, must be considered in light of the fact that the "best" method of determination of DIF is not obvious from these results. A clear elucidation of the "best" method of DIF determination should be performed with manufactured data in a Monte Carlo study. Such a study would eliminate any of the design aspects of this assessment that confound our results.

Causes

Two aspects of the assessment used in this study are potential causes of the unexpected results that we obtained. The assumptions used in the two techniques, ICC and LDF, are quite different and are differentially violated by this assessment.

Also, the problem of accounting for nonuniform DIF exists in the ICC method but not in the LDF method.

The ICC method is based upon item response theory. The assumptions of IRT include the concepts of unidimensionality, local independence and the invariance property of item characteristic curves in the population (Hambleton & Swaminathan, 1985). The LDF method is based upon the assumptions of discriminant analysis. The key assumption of discriminant analysis is that the variables and dispersion and covariance matrices are normally distributed. This technique is quite sensitive to violation of this assumption (Hair, et al., 1995).

The assessment instrument used in this project is analyzed for possible violations of the assumptions underlying each of these methods and theories. The most glaring problem with this assessment is that it most likely measures more than one construct, thus violating the assumption of unidimensionality. As is pointed out in the methods section of this report, five of the thirty items are designed to assess science and reading concurrently and seven of the items are designed to assess science and math concurrently. The remaining eighteen items are "pure" science items. Clearly, this assessment is not unidimensional.

There is additional evidence for the multidimensionality of this measure. We pointed out in the methods section of this paper that there are six domains of science upon which the assessment draws. These domains include concepts of science, nature of science, habits of the mind, attitudes, science processes and applications of science. Different items in the assessment emphasize varying combinations of these constructs. Once again, this assessment is not unidimensional.

The difficulties with multidimensionality in item response theory have been recently addressed by Douglas, Roussos, and Stout, (1996). In their 1996 article in the Journal of Educational Measurement, they discuss the multidimensionality assumption that is such a problem in this measurement instrument. They propose suggestions to resolve

this problem. It is beyond the scope of this project to attempt multidimensional IRT analysis, however, future research on this data set should include the consideration of these suggestions.

Another problem with this assessment is that it also violates the item response theory assumption of local independence (Hambleton & Swaminathan, 1985). For example, items # 1 through # 8 and # 18 all pertain to soil tests. Items #9 through #17 are concerned with levers. Items #19 through #30 all relate to salinity. When items are related to one another like this, they cannot actually be considered separate items in the item response theory sense. Again, Douglas, Roussos and Stout address this issue in their 1996 article. They suggest the analysis of item "bundles" or "testlets" as a technique for overcoming this difficulty. In this assessment, for instance, we should use their suggestions to analyze the "soil test", "levers" and "salinity" testlets. It is beyond the scope of this project to do such an analysis. However, once again, it would be an appropriate direction for additional research.

An additional problem with the design of this assessment is that the students work independently at times and, at other times, work in pairs or groups of four. The measures of individual ability that are produced by an item response theory analysis of this data are highly suspect. These measures reflect some individual ability and some collaborative ability or even "acquired" ability. It has been our experience that students working in groups often produce work that is most reflective of the ability of those with the most profound understanding of the material and processes. It might be possible to treat group or pair collaborative items as test "bundles", as suggested by Douglas, Roussos and Stout (1996). This analysis could be conducted based upon the same assumptions as those for items that are connected on the basis of content. However, the measures of individual ability that might be produced with this technique are still completely questionable.

Item response theory analysis of this assessment is limited by three different

difficulties. The IRT assumptions of unidimensionality and local independence are violated in this assessment. Also, the group collaboration on many of the items confounds the measurement of individual ability. Techniques that might overcome the first two limitations have been proposed but are beyond the scope of this project.

The sensitivity of LDF to non-normality of the variables and variances can be a significant problem in this analysis. However, normality is no more or less likely with this performance assessment than it would be with more traditional multiple-choice tests. Consequently, the design of this indicator of performance is not responsible for the possible violation of this assumption nor for the differences in DIF determined by the two different methods.

It is clear then that a potential source of the discrepancy in DIF observed when comparing the ICC and the LDF methods could be due to flaws in the design of the assessment instrument upon which this analysis is based. The violations of the assumptions of IRT that undergird the ICC method are severe. On the other hand, the assumptions of discriminant function analysis, upon which LDF is based, are not probably violated in this sample. We can conclude that the differences in the results of the two methods could be due to the "violation of assumption" factor and that, given this situation, the LDF method is more likely to be accurate.

In addition to the violation of assumption issue, another possible reason for the differences in the results of the ICC method versus the LDF method may be explained by consideration of the uniformity of the DIF in the items that each method identified as functioning differentially. In the ICC method, the formula used to calculate the DIF assumed an unsigned area between the ICC curves. The results of this study indicate that two of the items with large DIF indices are in fact signed or nonuniform. In the LDF method, the formulas used to calculate the probability of membership in one of the gender groups take into consideration both signed and unsigned or nonuniform and uniform DIF. It must be concluded, again, based upon the "uniformity" factor, that the

LDF method is superior to the ICC method, in this situation.

Each of the two factors considered above could be responsible for the discrepant results between the two methods of DIF calculation (ICC and LDF). Both of these aspects of the determination of DIF are potential causes of the unanticipated results from this study. The violations of assumptions and the uniformity of the DIF are possible sources of error in the determination of DIF. In both cases, however, the errors produced by the problems are more likely to affect the ICC method than the LDF method.

Comparisons

LDF

In addition to the possible causes of the unexpected results of our method comparison in this study, there are several other factors in regard to each of these methods that differentially recommend one of these techniques over the other. LDF uses simultaneous determination of response categories, produces a significance measure for DIF using the confidence intervals around the "item plus total" scores and is easier to use than the ICC method. The ICC method is not sample dependent and it provides model specificity that the LDF method does not.

Logistic Discriminant Function analysis of DIF has three advantages over Item Characteristic Curve analysis of DIF. The first of these advantages is that LDF simultaneously determines DIF for all of the multiple response categories. The ICC method calculates an index for the first pair of responses, then the second pair and so on until all pairs of responses have been considered. Simultaneous determination is a considerable advantage over pairwise determination. LDF determines the DIF of all the responses to a particular item at the same time.

Another advantage of the LDF method over the ICC method is that it produces a significance measure to which the DIF can be compared. Significance measures can

be implemented with the ICC method of determination of DIF. However, the .95 confidence interval determination with the LDF method is easier and an expected part of the analysis.

The third advantage of the LDF method is that it is easier to use than the ICC method. It is less computer and operator intensive. Its calculation can be performed on SPSS, a more common software package for statistical analysis than Multilog (SSI, 1991). Thus, there are three advantages to the use of the LDF technique in the determination of differential item functioning. It is easier to use, produces an significance measure and simultaneously determines the DIF for all responses to an item.

ICC

The ICC method of DIF determination has advantages as well. In particular, IRT techniques are not dependent upon the sample and provide model specificity for DIF analysis. Because the ICC method of DIF determination relies on the techniques of IRT, it is not sample dependent.

The fact that IRT models can specify a parameter for the discrimination of items (a) and a lower asymptote for guessing (c) is also an advantage when using the ICC method of analysis for DIF. However, the assessment used in this study can not take full advantage of this model specificity. The assessment upon which this study is based is a performance assessment. Guessing is not as much of an issue in performance assessments that require the demonstration of knowledge as it is in more traditional multiple-choice tests (Camilli & Shepard, 1994). Consequently, this particular assessment instrument does not take advantage of this strength of the IRT model and its ability to set a lower asymptote for guessing. The discrimination parameter that is possible to include in a two-parameter IRT model, can increase the model specificity of the DIF calculations in this measurement instrument. However, the advantage of being able to specify item discrimination does not necessarily outweigh

the inability to make use of the capacity to specify a lower asymptote.

Thus, the IRT advantages of model specificity and sample independence are not fully realized in the assessment used in this study. On the other hand, the advantages of the LDF technique are all clearly operational in this assessment. Those advantages include the simultaneous calculation of multiple responses, a significance test for the existence of DIF and the ease of use. Also, the issues of violation of assumptions and uniformity of DIF support the choice of the LDF method for this particular assessment. Clearly, in the case of this assessment instrument, the LDF method is superior to the ICC method for the calculation of DIF.

The previous analysis of the advantages and disadvantages of each of these two methods (ICC and LDF) for the determination of DIF in a performance assessment provides a useful template for test analysts to use when deciding which techniques to employ when conducting differential item functioning analyses. Such analysis provides evidence for the feasibility of using a simultaneous estimation procedure for polytomously scored items. In that context, it may improve the detection of actual DIF in performance assessments that rely upon such items. Considering the fact that such assessments are being suggested as more "authentic" than traditional tests and are proliferating (Wiggins, 1989), determination of differential item functioning and possibly test bias for these instruments is of paramount concern.

TABLE 1:

Indices of DIF as well as the a and b parameter estimates by gender for 30 items using the Lee methodology-State Performance Assessment Program

item #		a's	b-1's	b-2's	b-3's	b-1's DIF	b-2's DIF	b-3's DIF
1	girls	0.91	-1.25			0.16		
	boys	0.98	-0.85					
	difference	0.07	-0.4					
2	girls	1.02	-0.56			0.22		
	boys	1.17	-0.05					
	difference	0.15	-0.51					
3	girls	1.24	-0.11	1.38		0.27	0.19	
	boys	1.42	0.45	1.78				
	difference	0.18	0.56	0.4				
4	girls	1.23	-0.14	1.88		0.26	0.16	
	boys	1.51	0.38	2.17				
	difference	0.28	0.52	0.29				
5	girls	1.19	-0.16	2.25		0.26	0.11	
	boys	1.42	0.37	2.43				
	difference	0.23	0.53	0.18				
6	girls	0.82	-0.37	2.36		0.2	0.1	
	boys	0.91	0.14	2.59				
	difference	0.09	0.51	0.23				
7	girls	1.18	0.13	4.03		0.13	0.03	
	boys	1.24	0.41	4.14				
	difference	0.06	0.28	0.11				
8	girls	0.71	0.21			0.2		
	boys	0.83	0.72					
	difference	0.12	0.51					
9	girls	0.7	-0.04			0.13		
	boys	0.8	0.29					
	difference	0.1	0.33					
10	girls	1.11	1.45			0.07		
	boys	1.31	1.4					
	difference	0.2	-0.05					
11	girls	0.94	1.21			0.12		
	boys	1.25	1.21					
	difference	0.31	0					

item #		a's	b-1's	b-2's	b-3's	b-1's DIF	b-2's DIF	b-3's DIF
12	girls	1.17	1.65			0.08		
	boys	1.27	1.48					
	difference	0.1	-0.17					
13	girls	0.69	1.44	2.46		0.03	0.03	
	boys	0.75	1.4	2.36				
	difference	0.06	-0.04	-0.1				
14	girls	0.74	-2.33	1.31	2.13	0.14	0.06	0.05
	boys	0.86	-1.85	1.3	2.07			
	difference	0.12	-0.48	-0.01	-0.06			
15	girls	0.87	0.7			0.12		
	boys	1.04	0.92					
	difference	0.17	0.22					
16	girls	1.06	0.38			0.14		
	boys	1.3	0.63					
	difference	0.24	0.25					
17	girls	0.97	-0.36	0.77		0.17	0.12	
	boys	1.27	-0.05	0.85				
	difference	0.3	-0.31	0.08				
18	girls	0.94	-3.2	0.74		0.32	0.11	
	boys	1.11	-2.19	0.95				
	difference	0.17	-1.01	0.21				
19	girls	0.56	-8.02	3		0.13	0.17	
	boys	0.91	-4.5	-4.5				
	difference	0.35	-3.52	-0.69				
20	girls	0.54	-7.89	3.28		0.15	0.15	
	boys	0.88	-4.38	-4.38				
	difference	0.34	-3.51	-0.76				
21	girls	0.87	1.85	4.22		0.09	0.02	
	boys	1.04	1.96	1.96				
	difference	0.17	0.11	-0.25				
22	girls	0.67	-2.1			0.3		
	boys	0.94	-1.19					
	difference	0.27	-0.91					
23	girls	0.5	-2	5.64		0.33	0.05	
	boys	0.75	-0.84	4.41				
	difference	0.25	-1.16	-1.23				

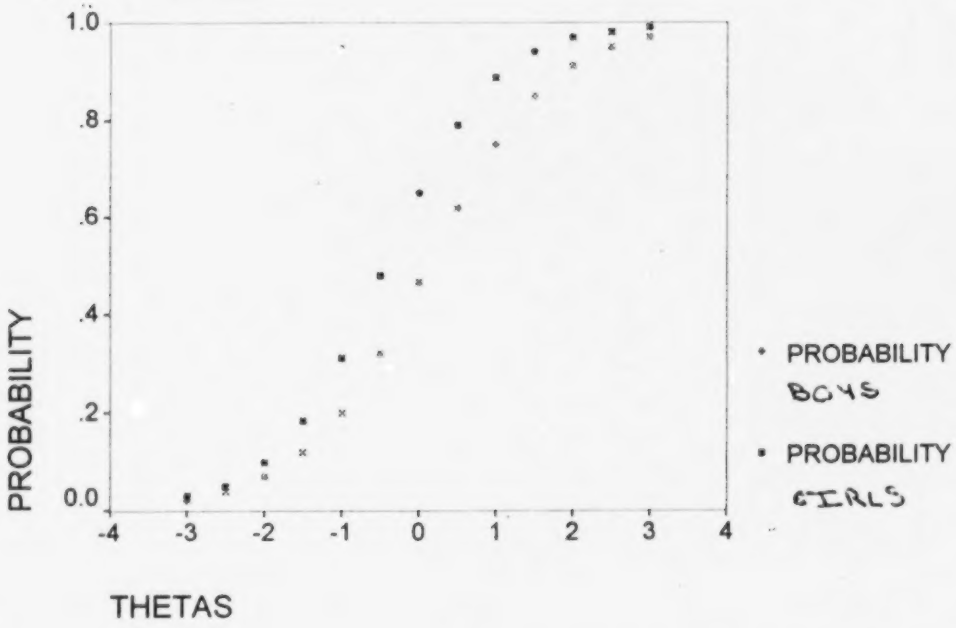
item #		a's	b-1's	b-2's	b-3's	b-1's DIF	b-2's DIF	b-3's DIF
24	girls	0.64	1.23			0.15		
	boys	0.9	1.05					
	difference	0.26	-0.18					
25	girls	1.01	-1.75	2.42		0.31	0.11	
	boys	1.32	-1.05	2.16				
	difference	0.31	-0.7	-0.26				
26	girls	0.58	-1.82	4.04		0.32	0.09	
	boys	0.85	-0.81	3.29				
	difference	0.27	-1.01	-0.75				
27	girls	0.91	0.14	3.74		0.21	0.06	
	boys	1.19	0.56	3.31				
	difference	0.28	0.42	-0.43				
28	girls	0.92	-1	4.84		0.22	0.08	
	boys	1.2	-0.52	3.92				
	difference	0.28	-0.48	-0.92				
29	girls	1.29	0.09	1.45		0.1	0.07	
	boys	1.57	0.21	1.41				
	difference	0.28	0.12	-0.04				
30	girls	1.18	0.74	3.4		0.12	0.02	
	boys	1.36	0.97	3.27				
	difference	0.18	0.23	-0.13				

Appendix

A

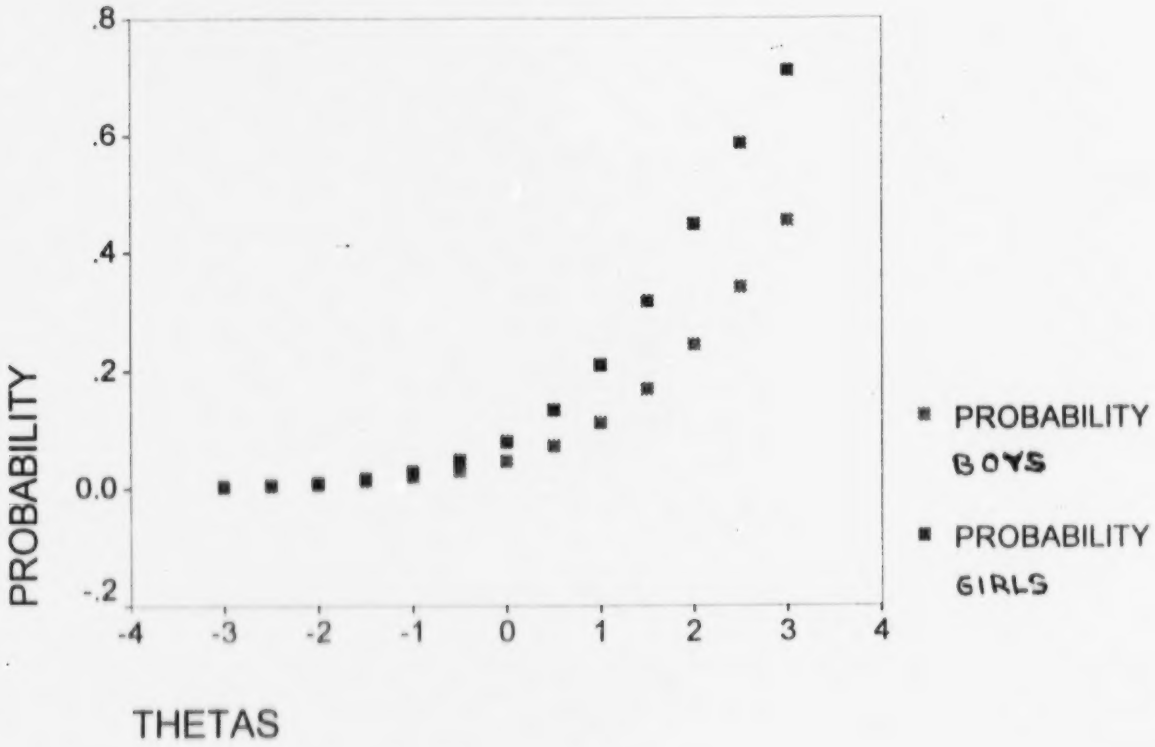
ICC FOR ITEM 3

b1 dif .22



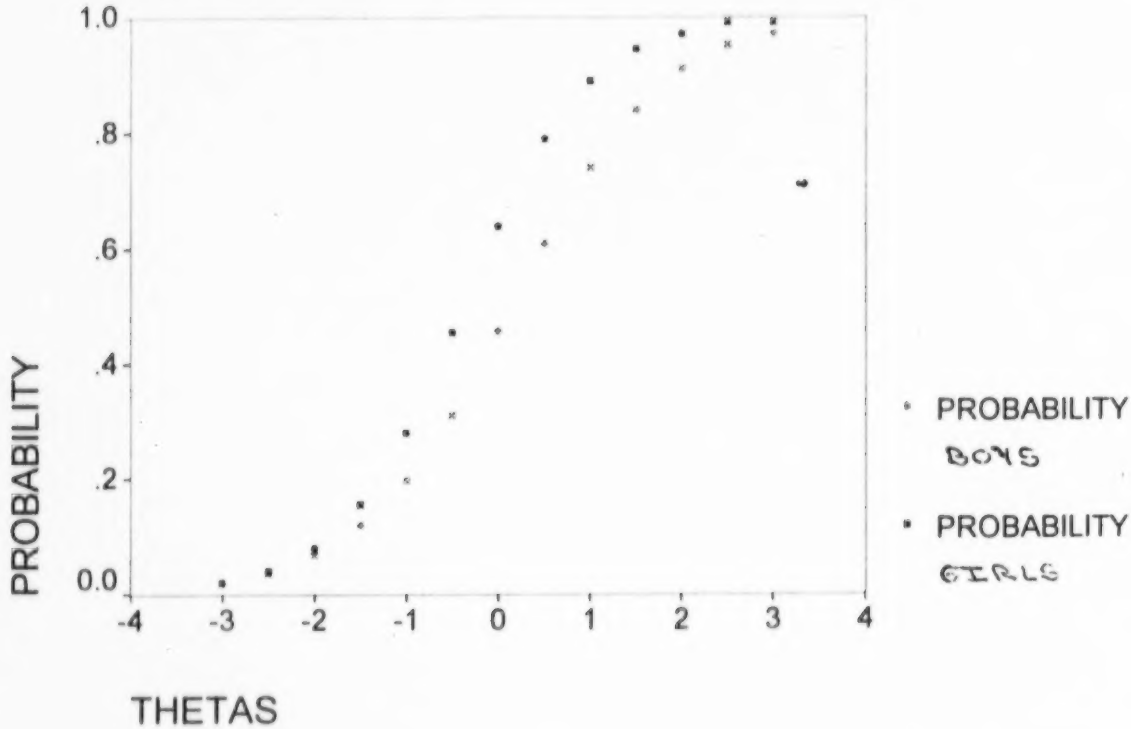
ICC ITEM 18

b1 dif .32



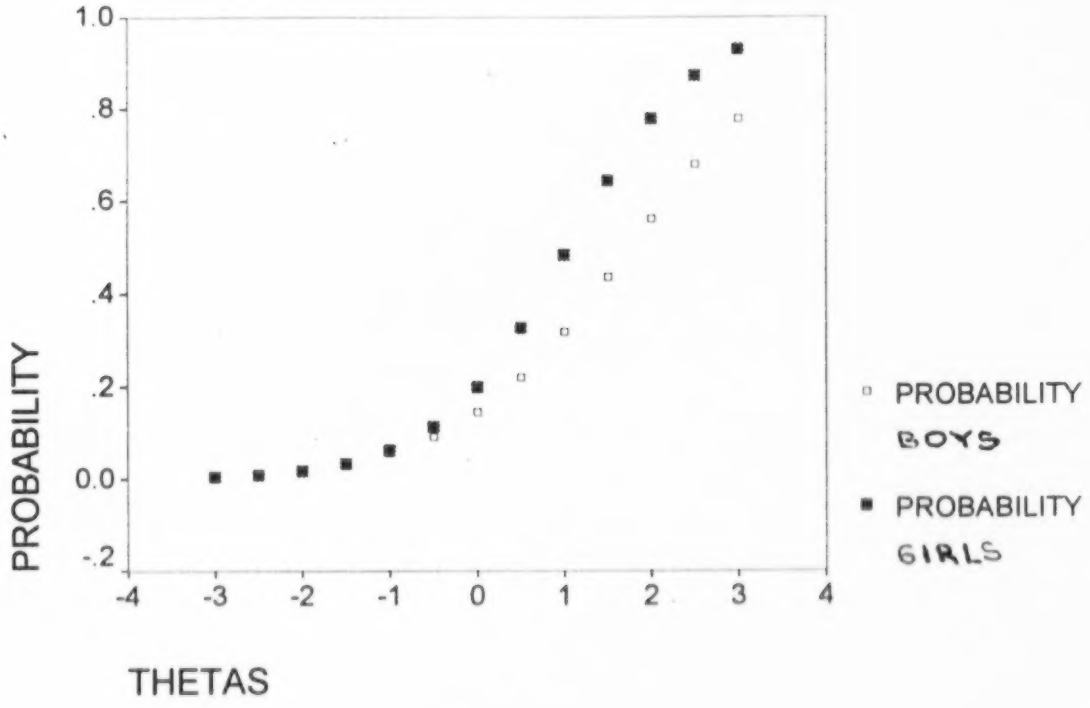
ICC FOR ITEM 4

b1 dif .26



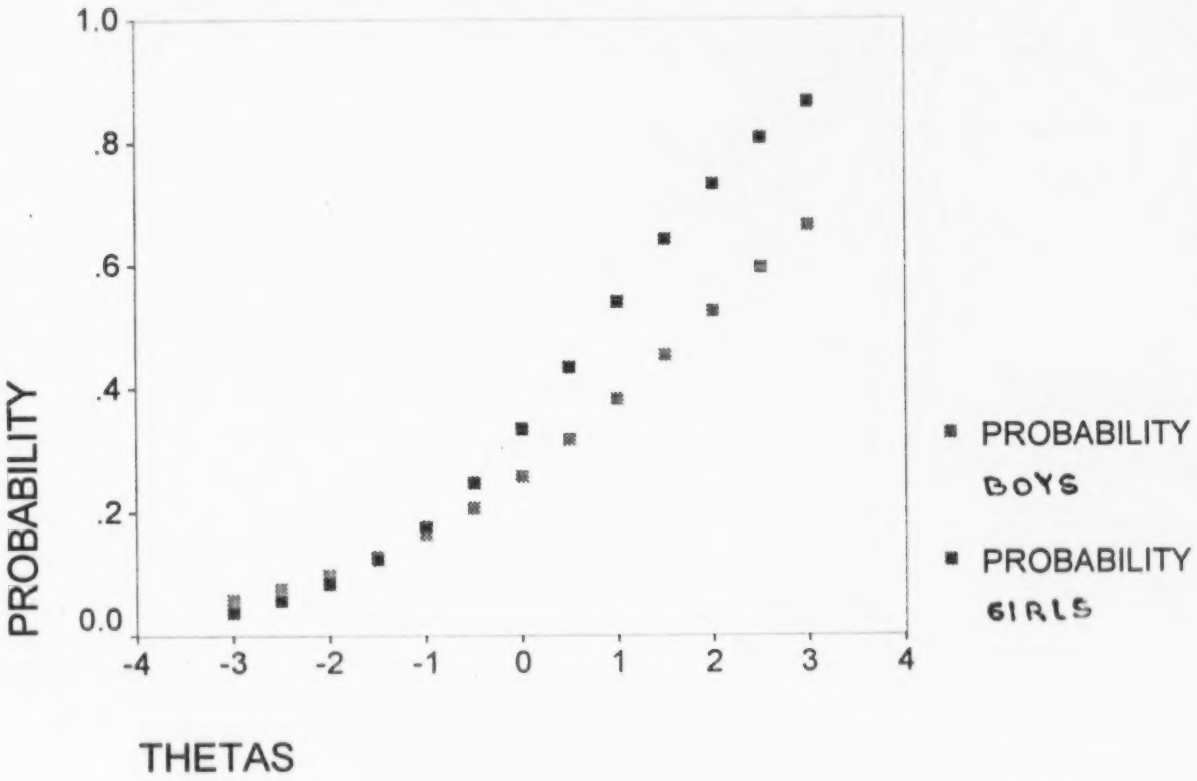
ICC ITEM 25

b1 dif .31



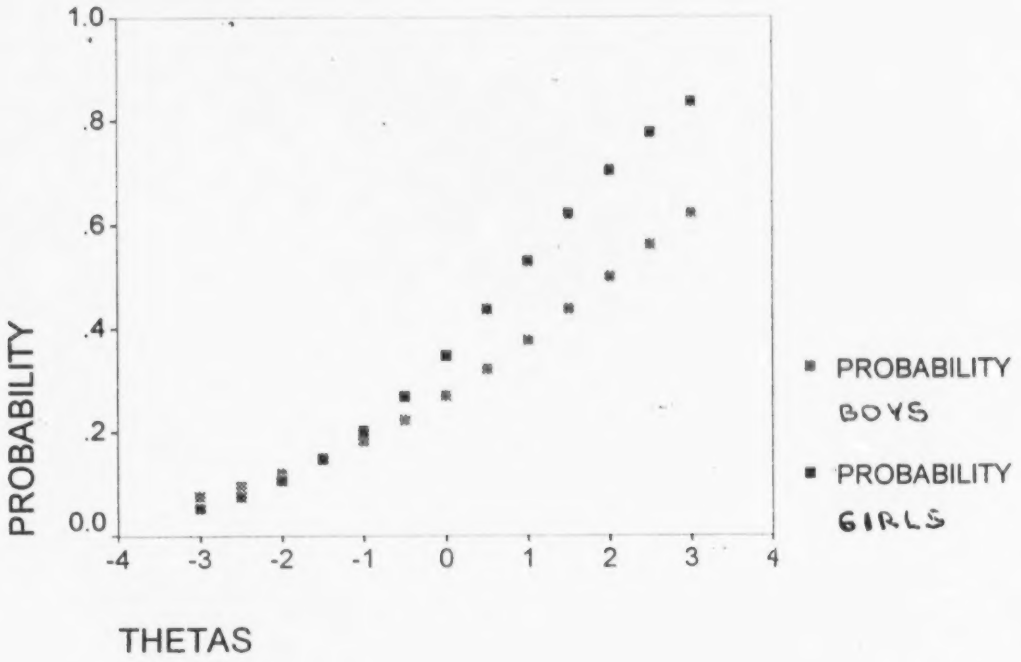
ICC ITEM 26

b1 dif .32



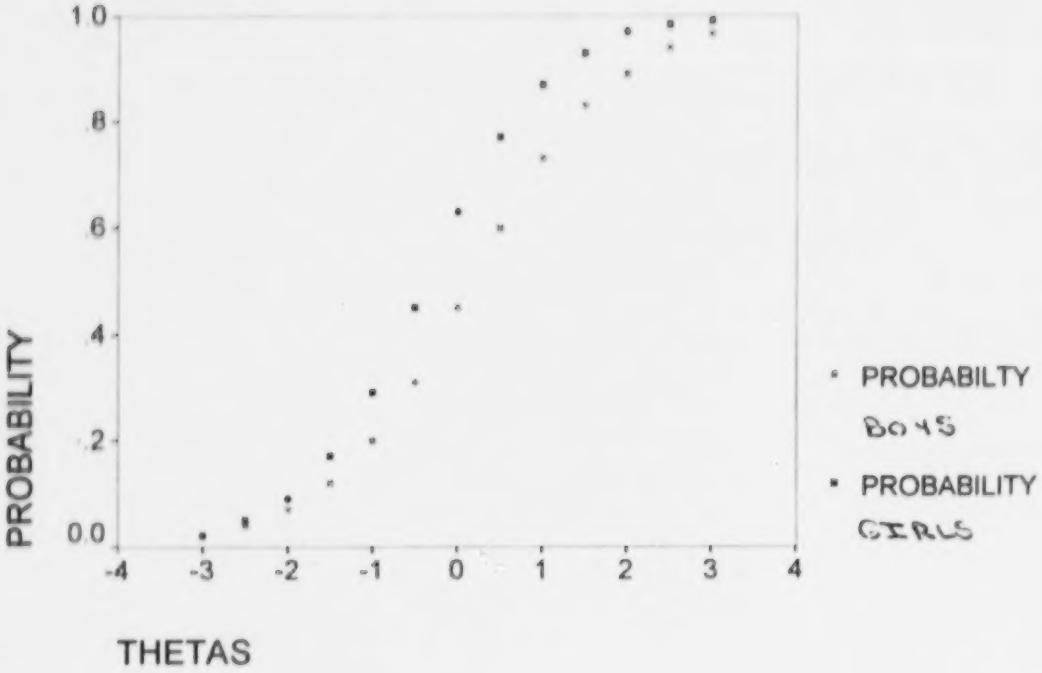
ICC ITEM 23

b1 dif .33



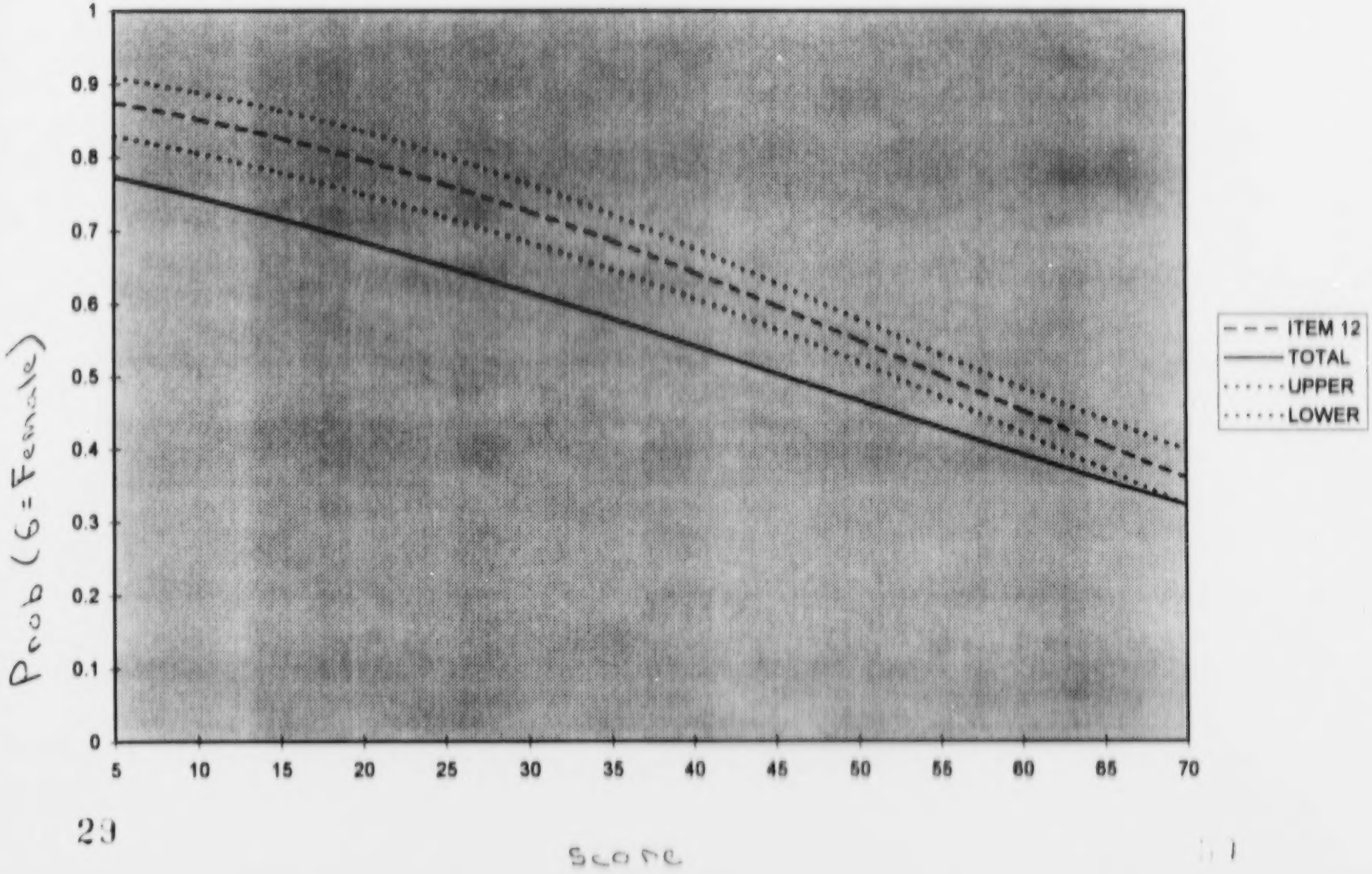
ICC FOR ITEM 5

b1 dif .26

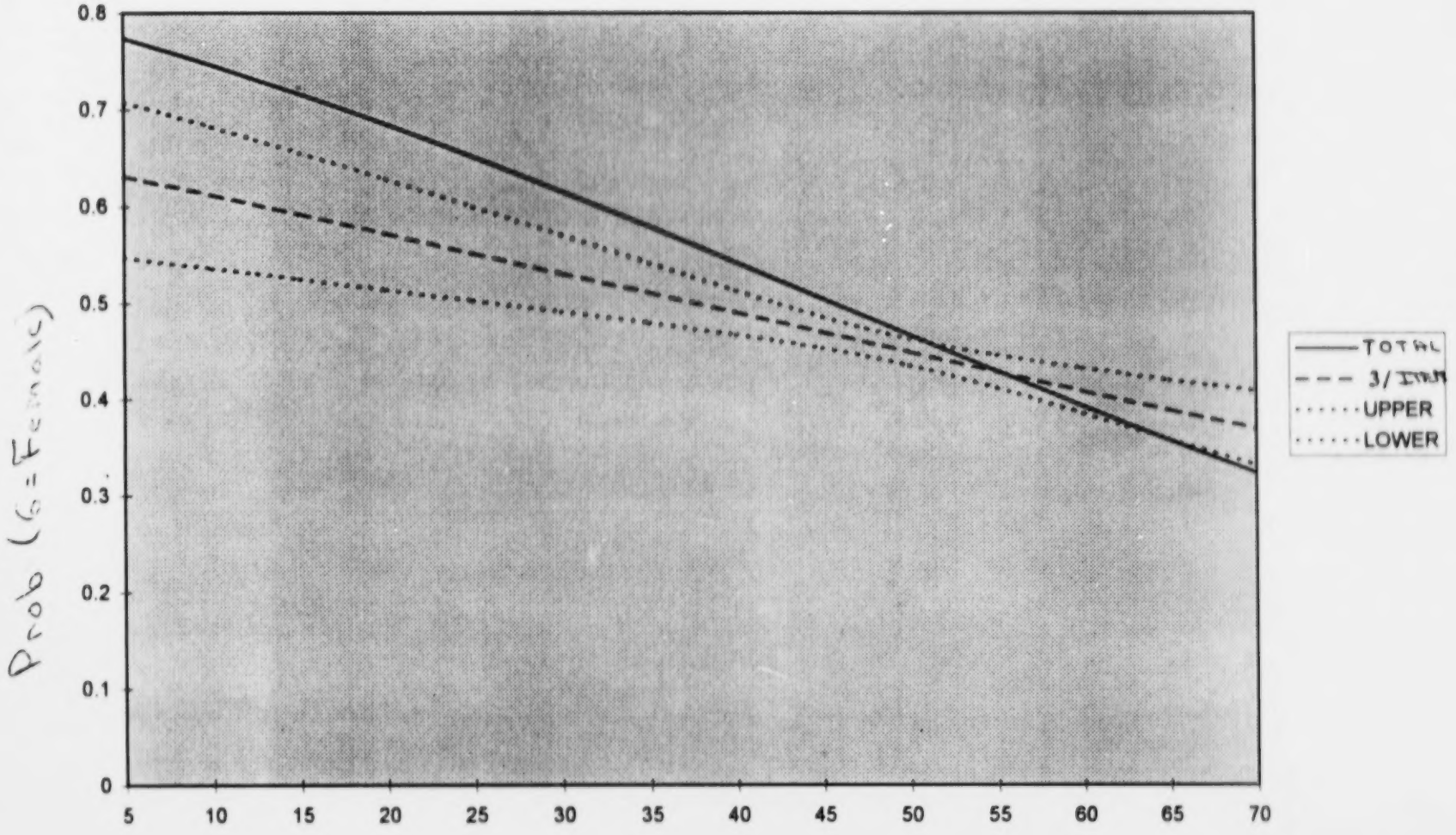


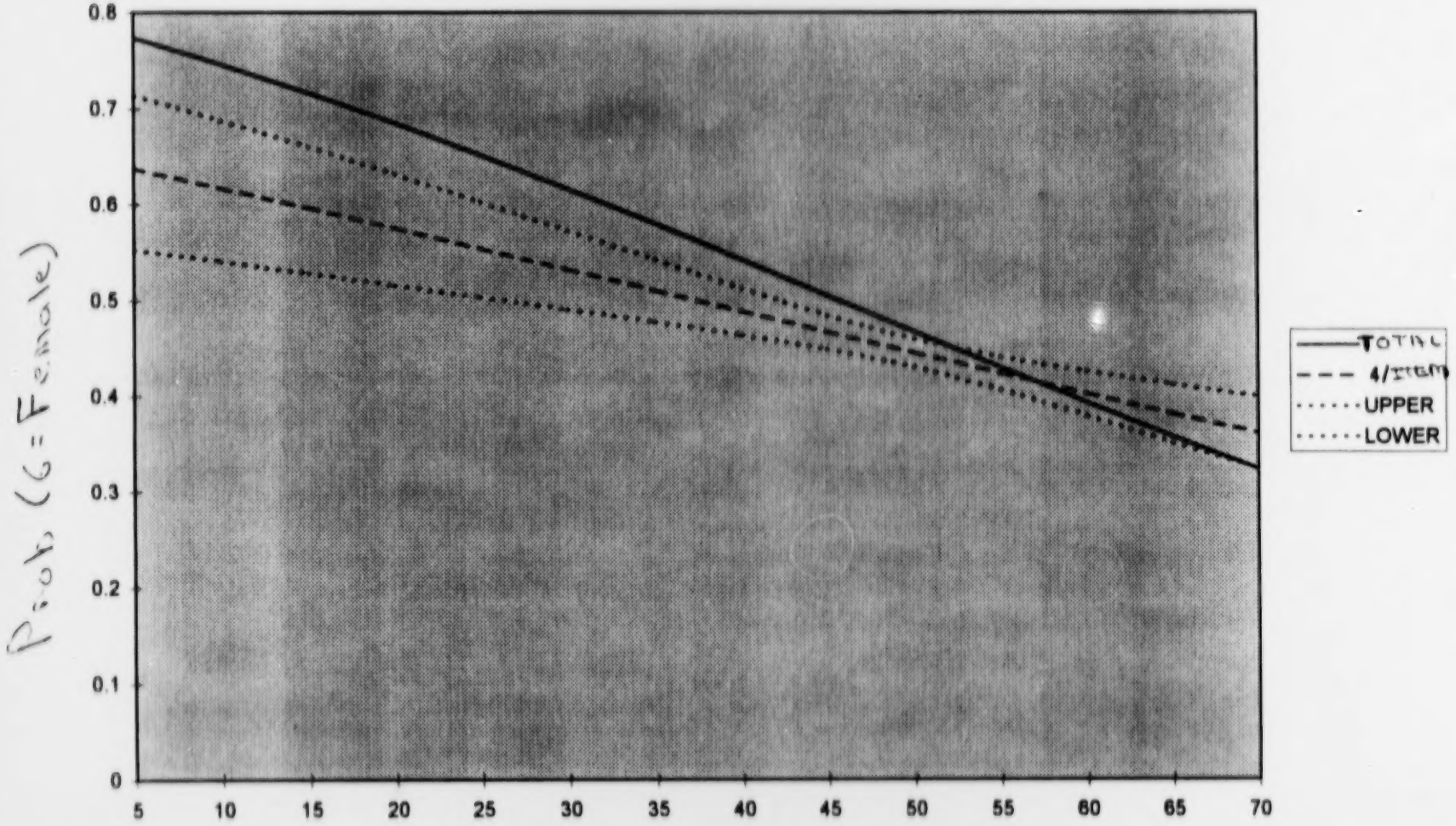
Appendix

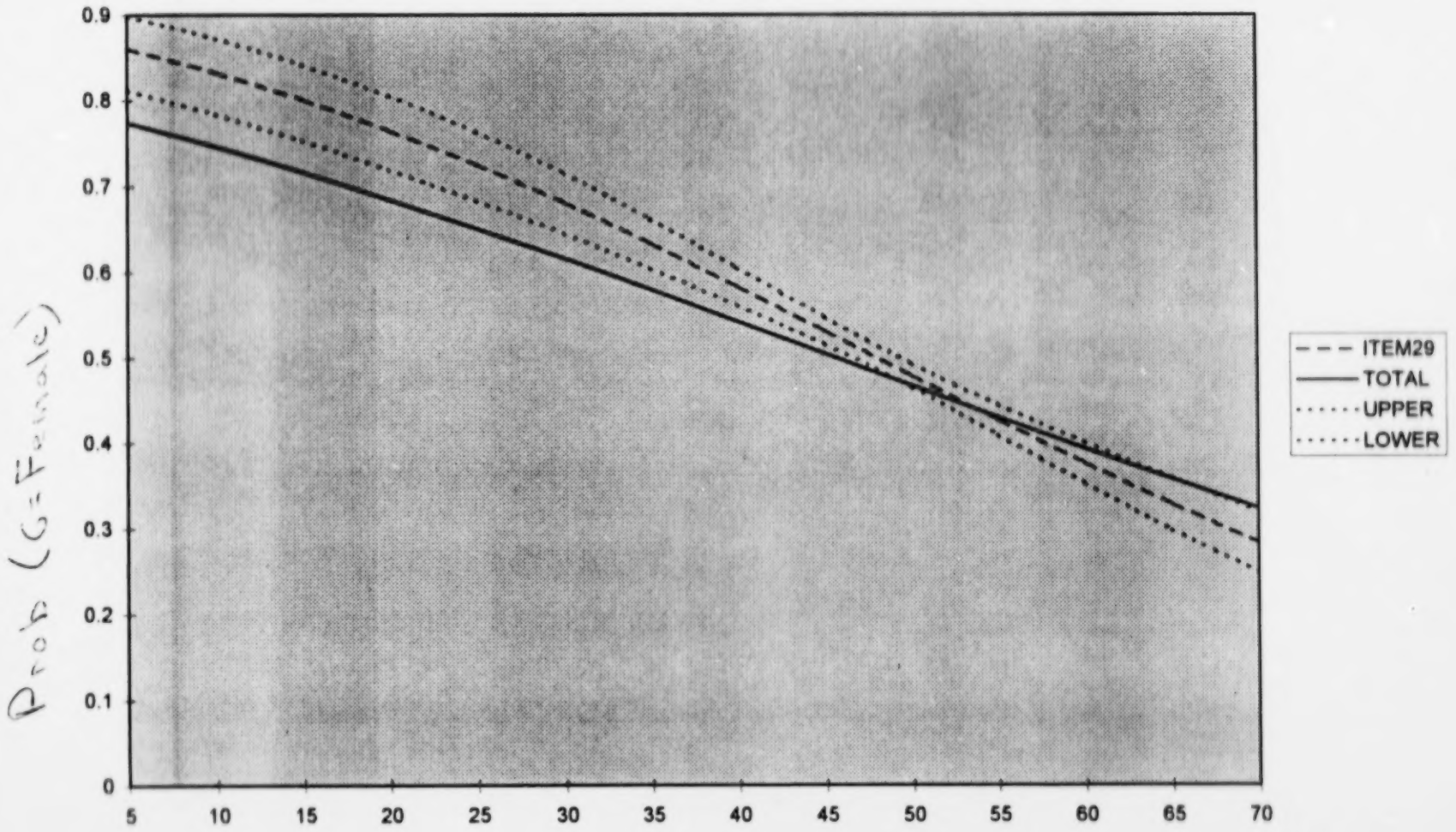
B

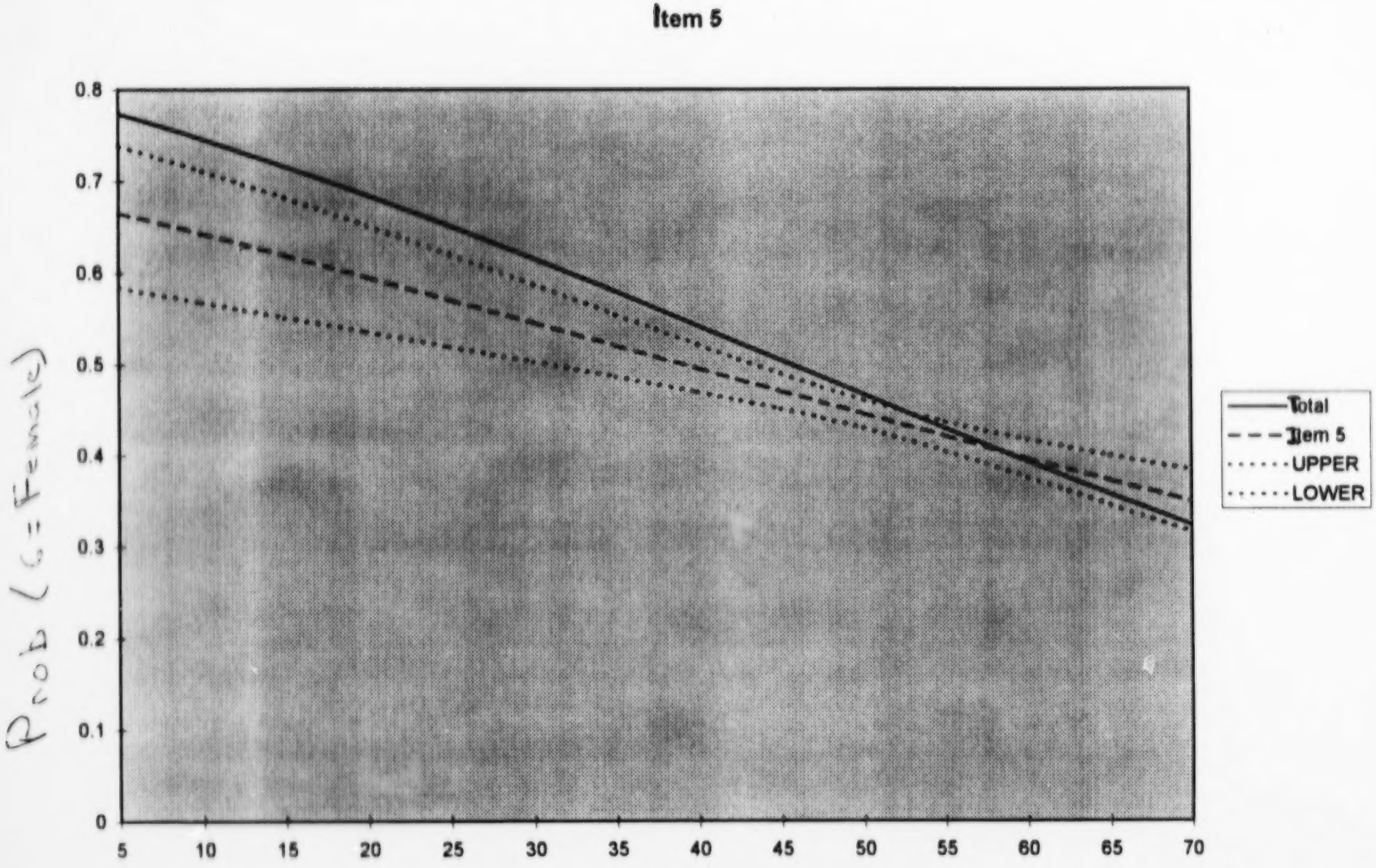

ITEM 12
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
Prob (G=Female)
Score
ITEM 12
TOTAL
UPPER
LOWER

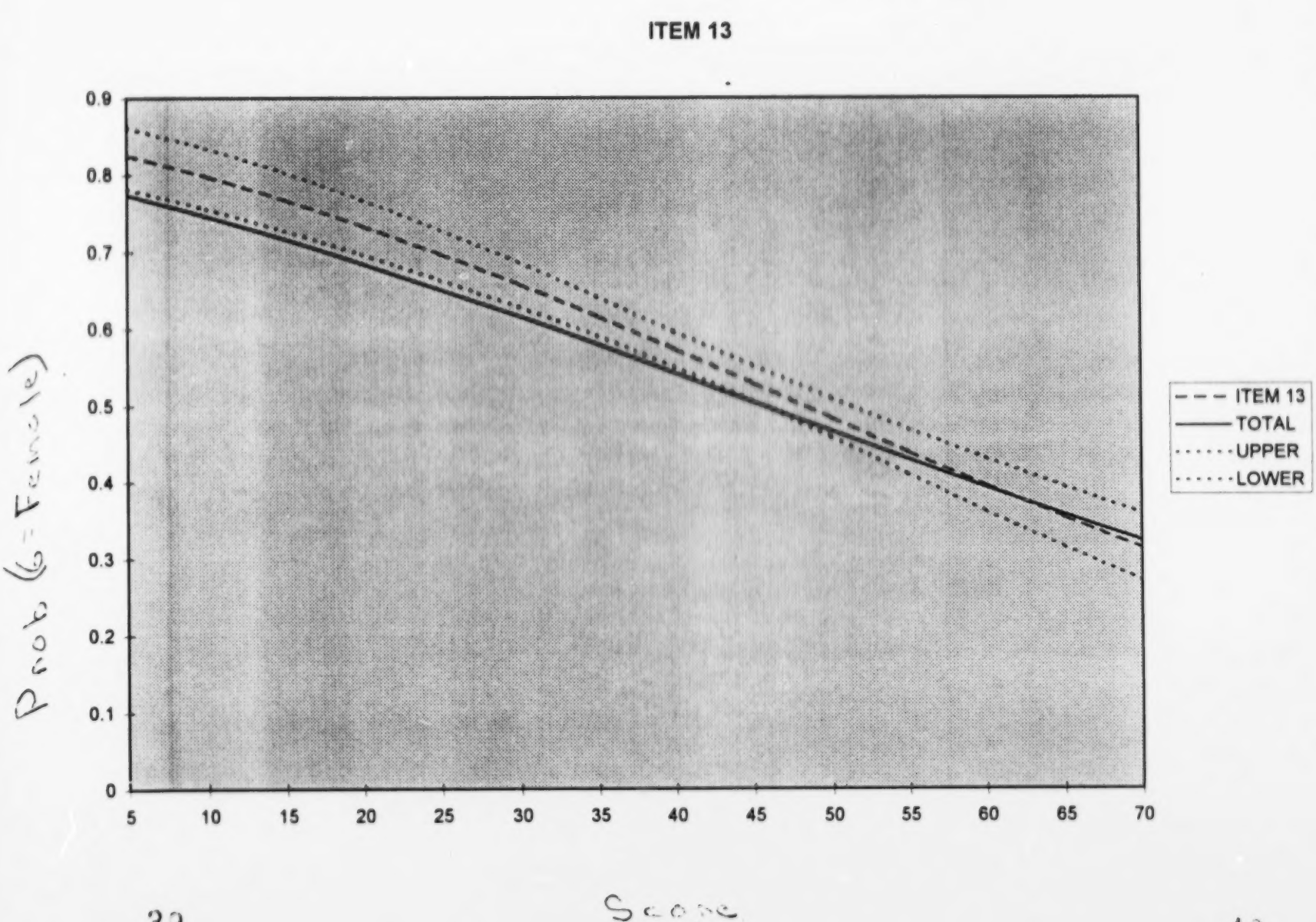
ITEM 3

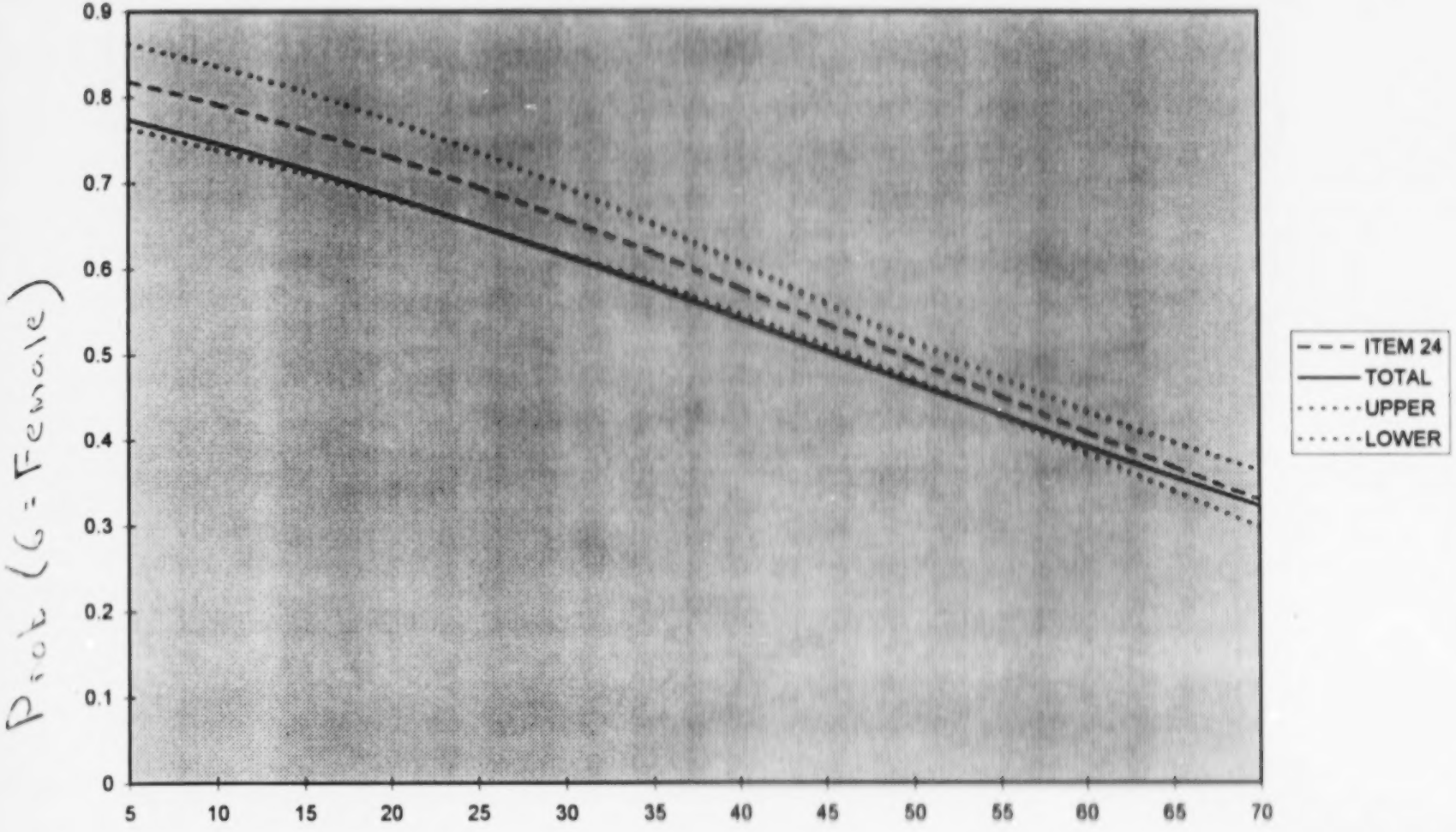

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
TOTAL
3/ITEM
UPPER
LOWER
Prob (G=Female)
Score


Item 4
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
TOTAL
4/ITEM
UPPER
LOWER
Prob (G=Female)
Score


ITEM 29
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
ITEM29
TOTAL
UPPER
LOWER
Prob (G=Female)
Score


Item 5
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
Total
Item 5
UPPER
LOWER
Prob (Correct)
Score


ITEM 13
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
ITEM 13
TOTAL
UPPER
LOWER
Prob (G=Female)
Score


ITEM 24
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
Prob (G = Female)
ITEM 24
TOTAL
UPPER
LOWER

41

Score

42

References

Berk, R. A. (Ed.). (1982). Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Camilli, G. & Shepard, L. A. (1994). Methods for identifying biased test items. Volume 4. Thousand Oaks, CA: Sage Publications.

Douglas, J. A., Roussos, L. A. & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. Journal of Educational Measurement, 33, (4), 465-484.

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). Multivariate data analysis with readings, 4th edition. Englewood Cliffs, NJ: Prentice Hall.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff Publishing.

Hauck, W. W. (1983). A note on confidence bands for the logistic response curve. The American Statistician, 37, 158-160.

Miller, T. R. & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. Journal of Educational Measurement, 30, (2), 107-122.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

Thissen, D. (1991). Multilog User's Guide, Version 6.0. Chicago, IL: Scientific Software Inc..

Wiggins, G. (1989). A true test: Toward a more authentic and equitable assessment. Phi Delta Kappan, 70, 703-713.

Yen, W. M., et al. (1992). Final technical report: Maryland School Performance Assessment Program 1991. Monterey CA: CTB Macmillan/McGraw-Hill.

----- (1991). 386 - Multilog 6. Scientific Software, Inc.. Chicago, IL: Author.